Death of Disk Panel

Ted Wobber

MSR Silicon Valley

August 10, 2011

Disks over the Years

	Mid-1980s	2009	Improvement
Disk capacity	30 MB	500 GB	16667x
Maximum transfer rate	2 MB/s	100 MB/s	50x
Latency (seek + rotate)	20 ms	10 ms	2x
Capacity/bandwidth (large blocks)	15 s	5000 s	333x *worse*
Capacity/bandwidth (1KB blocks)	600 s	58 days	8333x *worse*
Jim Gray's Rule [11] (1KB blocks)	5 min.	30 hours	360x *worse*

Source: J. Ousterhout et al. , The Case for RAMClouds: Scalable High-Performance Storage Entirely in DRAM, *SIGOPS Operating Systems Review 43(4).*

Are Disks Really Dead?

- What are the other options?
  - Tape
  - SSDs
  - Big Memory (e.g., RAMCloud)
  - Phase-Change Memory
  - Spintronics (aka MRAM, Racetrack)

Tape

- 4 Terabytes (per cartridge) uncompressed
- Less than $.10 per Gbyte
- ~250 MByte/s bandwidth (uncompressed)
- Seek latency in seconds to minutes
- Power: 51 watts
- Cost: $43,000 (+ shuttle and media costs) = ~$200K

- Combined with shuttle: 900 PBytes

Disk

• 6-Gb/s SAS/SATA drives	$440	$220
• Capacity (GB):	600	2000
• Spin Speed (RPM):	15,000	7200
• Average latency (ms):	2.0	4.2
• Random read seek time (ms):	3.4	8.5
• Random write seek time (ms):	3.9	9.5
• I/O data transfer (sustained max):	204MB/s	150MB/s
• Unrecoverable read errors:	1 in 10^{16}	1 in 10^{15}
• Average idle power:	11.68W	5.69W
• Average operating power:	16.35W	9.57W

SSD

- 6 Gb SATA drive ~$550
- Capacity: 240 GB

- Sequential Read 510 MB/s
- Sequential Write 240 MB/s
- 4KB Random Read 58,500 IOPS (230 MB/s)
- 4KB Random Write 48,500 IOPS (190 MB/s)

- **Power** Idle: 1.65 Watts; Active: 3 Watts

SSDs (cont)

- NAND flash is a odd animal
  - No over-write (OS TRIM support important)
  - Erase at 64-256x granularity of write
  - Limited erase cycles (~3-5K for MLC, 100K for SLC)
  - Read disturb / write disturb
  - Retention varies inversely with wear
  - Error correction vs. scale
  - FTL idiosyncrasies (compaction, wear-leveling)
- SSD market is becoming quite specialized
- SLC disappearing at low end

Big Memory

- For example, RAMCloud (Ousterhout, et al.)
- Clusters of RAM; very low latency
- Example configuration* (2009 pricing):

  1000 servers @ 64 GB/server

  Capacity: 64 TB

  Total cost: $4M

  Cost/GB: $60

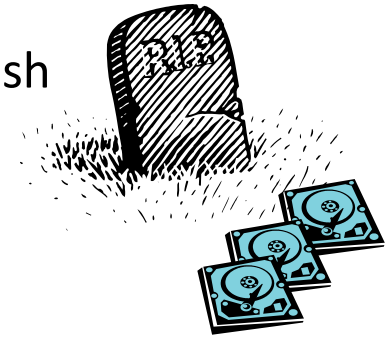
  Throughput: 10^9 ops/sec

* From: J. Ousterhout et al. , The Case for RAMClouds: Scalable High-Performance Storage Entirely in DRAM, *SIGOPS Operating Systems Review 43(4).*

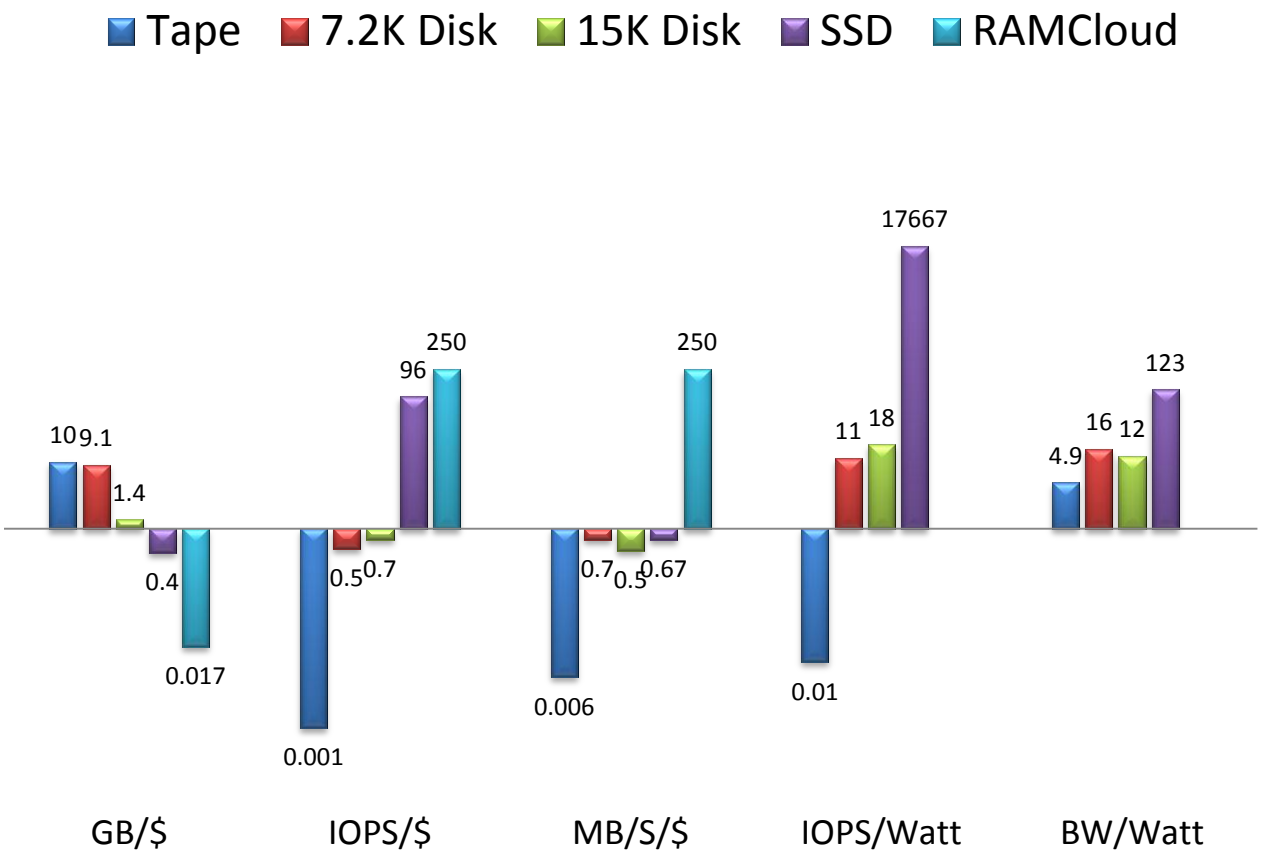
PCM + Spintronics

- Phase change memory
  - Resistance differences between crystalline and amorphous states
  - Factor of 10-100 in speed, and endurance compared to flash
  - Byte addressable
  - Thermal process: high current density; expansion/contraction border
  - 128Mb parts currently (at 90nm)

- Spintronics
  - Magnetic-resistive memory (e.g., MRAM, RaceTrack)
  - Very good scale, speed, and endurance compared to flash
  - Gigabit chips in 3-4 year at ~DRAM cost



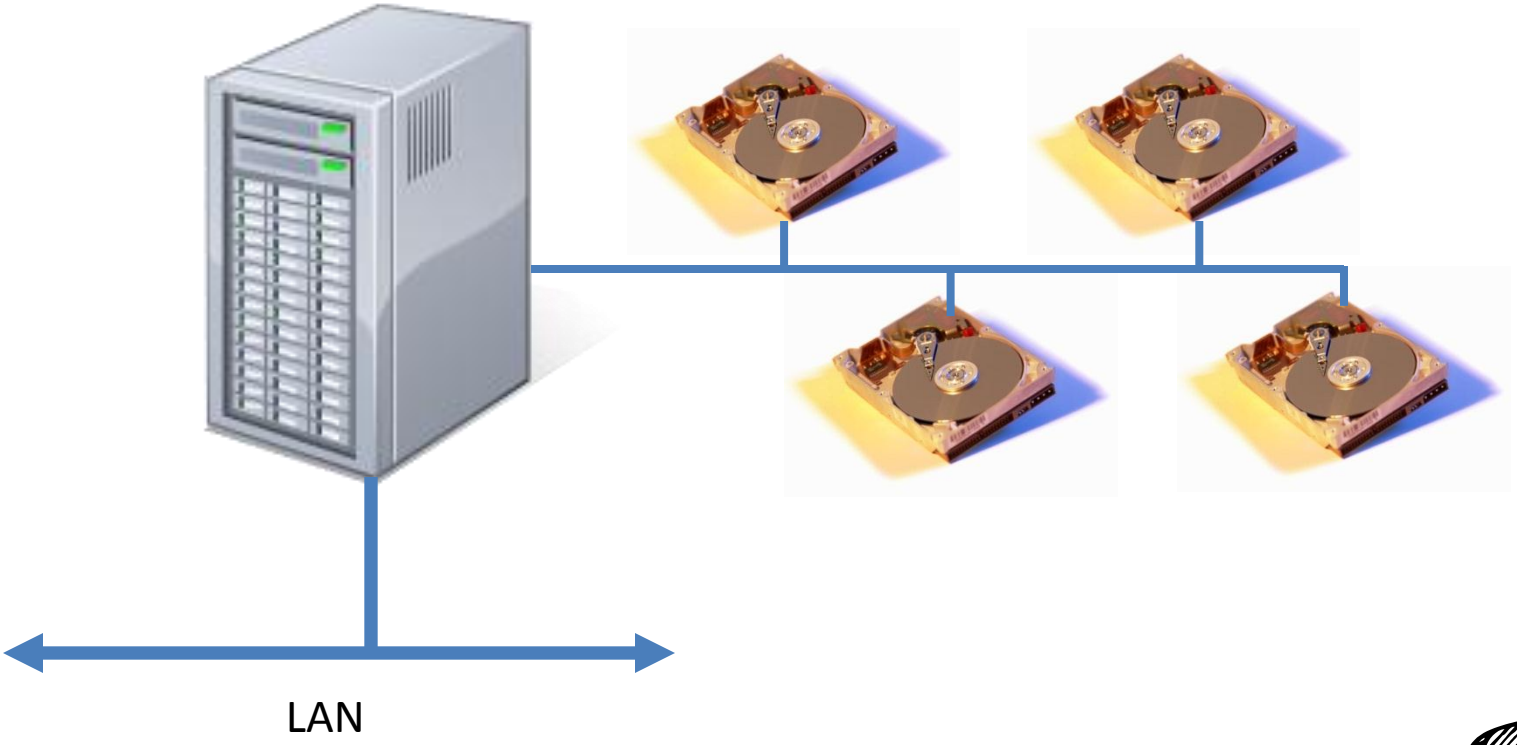
Some Comparisons



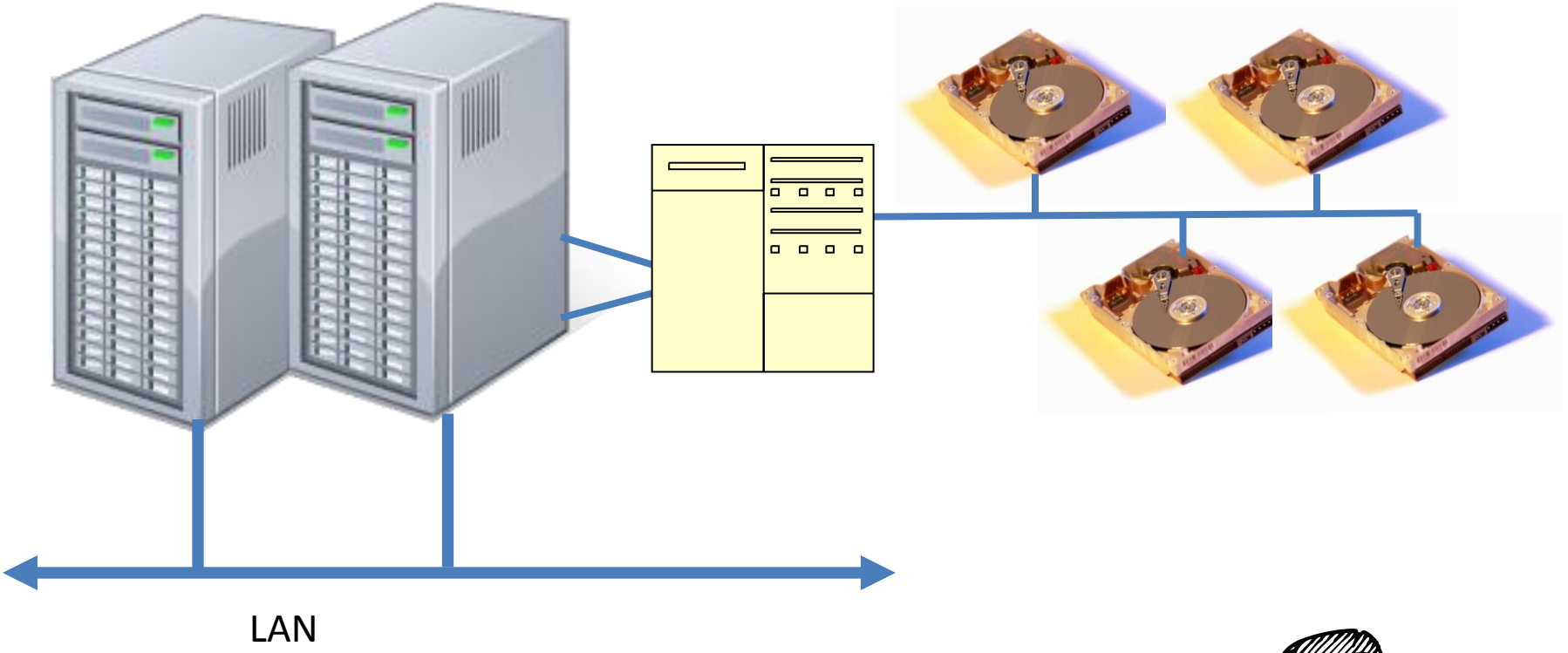
On the Merits

Tape	Disk	SSD	Big Memory
• Huge capacity • Offline storage • Streaming	• Cheap bandwidth with capacity • Sequential workloads	• IOPS (Metadata, swapping, caching) • Read-mostly workloads • Power	• Distributed transactions • Distributed strong consistency

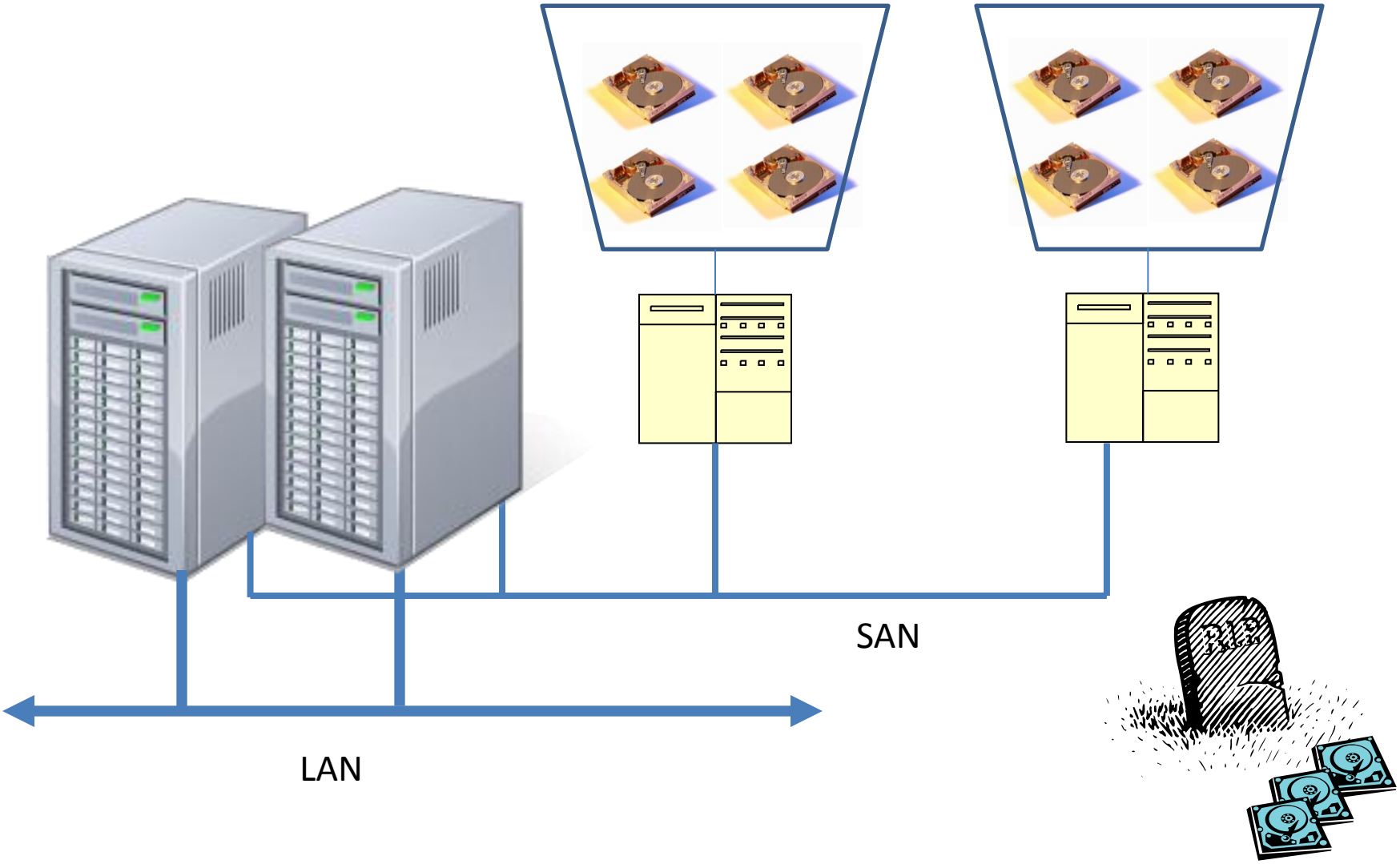
Servers and Disks



Servers and Storage Controllers

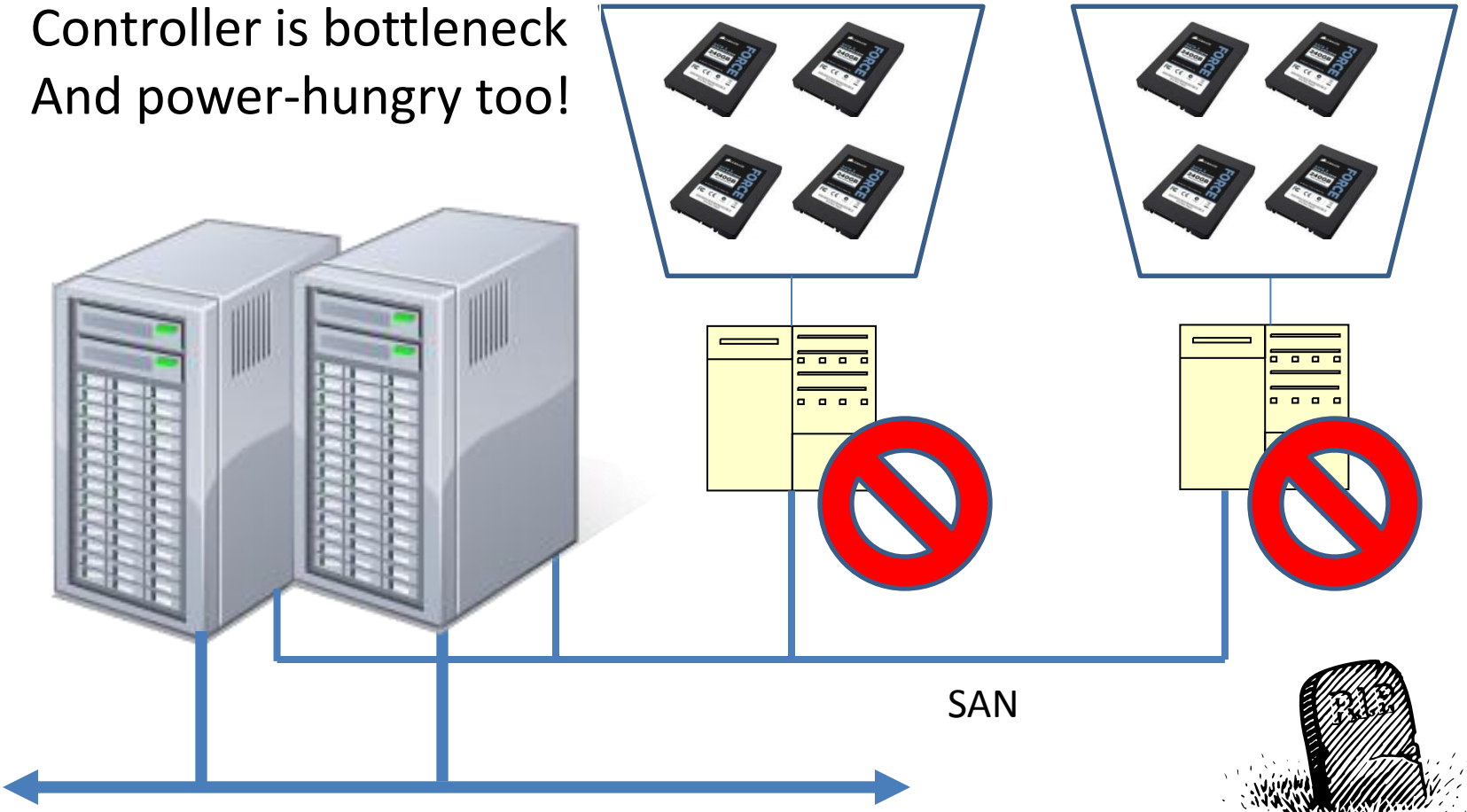


Servers and SCs and SANs



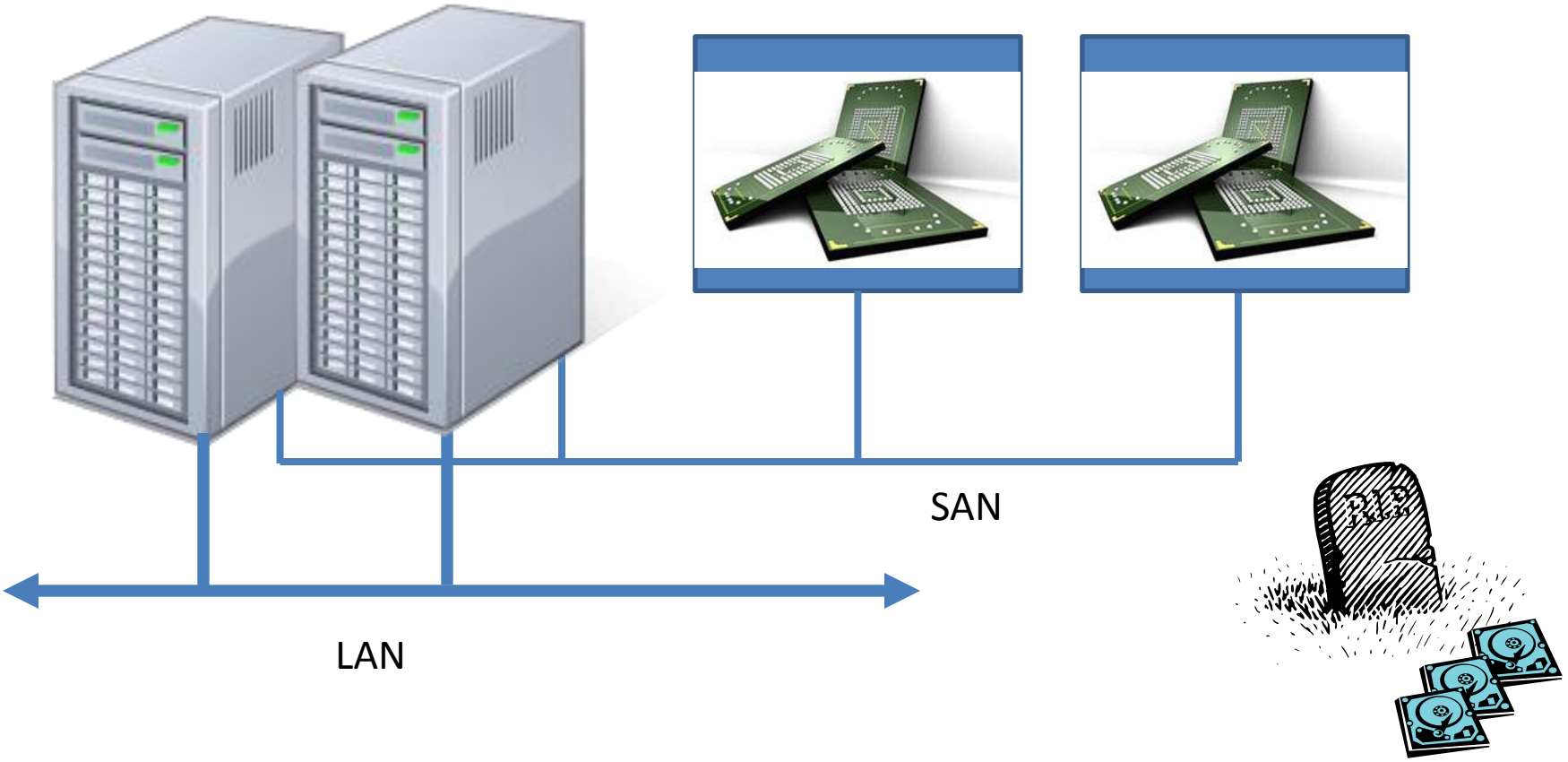
Servers and SCs and SSDs

- Controller is bottleneck
- And power-hungry too!



Servers and Flash Appliances

- Better power profile
- Well-tuned to flash
- SAN-interconnect is now bottleneck



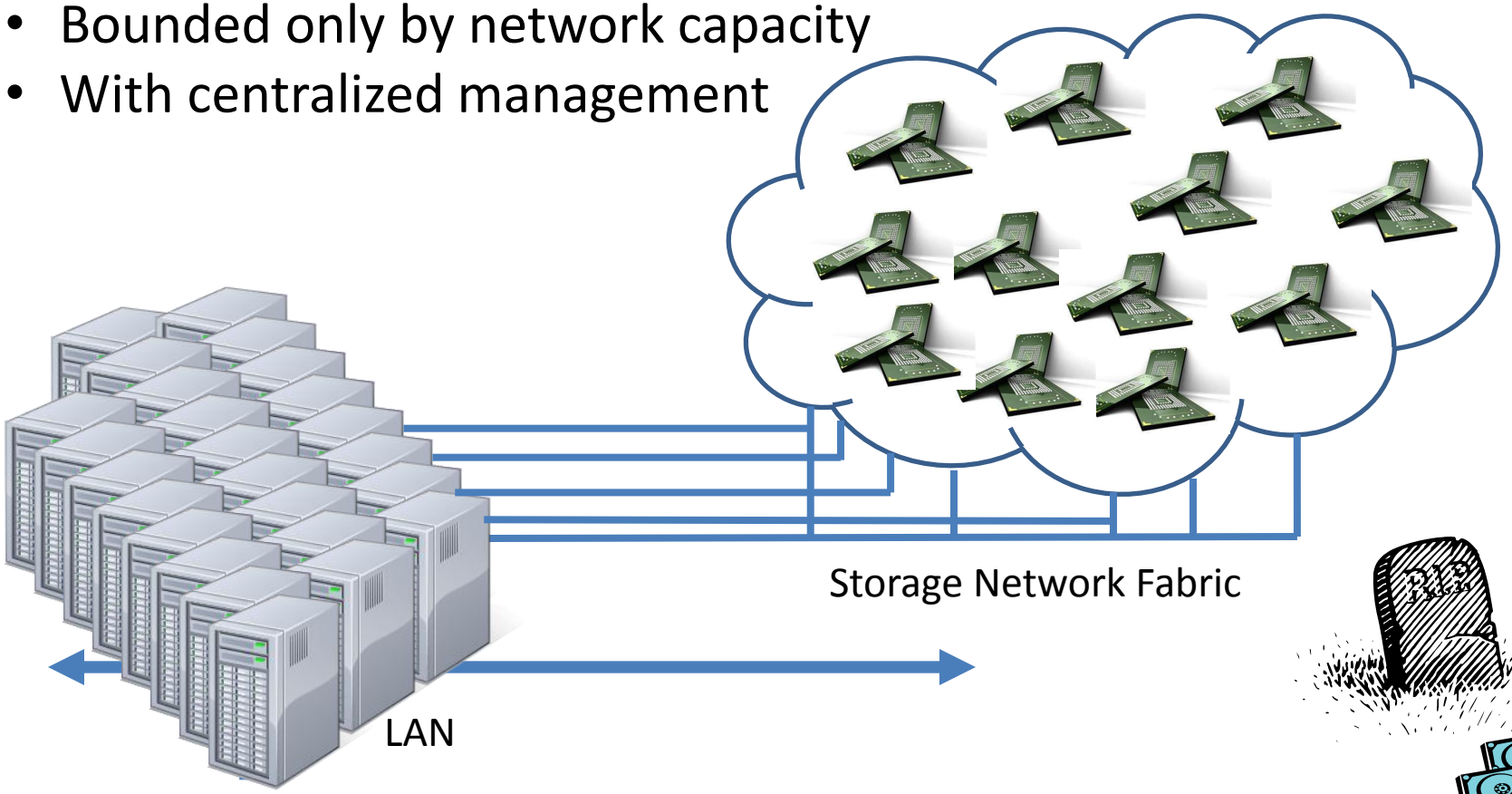
Do It in Parallel!

- Solid-state storage components have huge bandwidth / IOPS in aggregate
- Centralized storage controllers work hard to keep up
- Available BW / IOPS overwhelm single compute nodes
- How can we best distribute and consume these I/O resources?

Flash Clusters

(CORFU: **C**lusters **o**f **R**eplicated **F**lash **U**nits)

- Cluster of low-cost, low-power network attached flash
- Organized as a log to support distributed data consistency
- Bounded only by network capacity
- With centralized management



Is Disk Really Dead ?

- Replaced by Tape?
  - SERIOUSLY?: Tape has huge capacity, but high latency, high power consumption, fragile infrastructure, and high bandwidth cost
- Replaced by Flash?
  - NO: Power tradeoffs are nice, great IOP/s, but high cost per GB; scale-down difficulties; durability questions (especially for MLC)
- Replaced by other solid-state?
  - PROBABLY, but over time. Too soon to tell.
- Replaced by Big Memory?
  - NO: High memory cost, power, persistence.

Conclusion: No Surprises

- Evolutionary change is the rule
- Solid-state devices will slowly displace disk for many, but not all, things
- Solid-state devices will drive innovation with respect to interconnect